Análisis de sentimientos de startups usando un enfoque basado en léxicos

Sentiment analysis of startups using a lexical-based approach

[1]Roberto Carlos Fortich Mesa, [2]Michael Muñoz Guzmán, [3]María Alejandra Santís Puche

[1,2,3]Corporación Universitaria Antonio José de Sucre, Colombia

Recibido: 10/20/24 Aprobado: 02/02/25

Resumen

En la última década, la proliferación de startups ha experimentado un notable crecimiento, paralelo a los avances en innovación tecnológica y acceso a financiación en América Latina. Este estudio explora las percepciones de los fundadores y gerentes de startups respecto a las expectativas de los inversionistas mediante técnicas avanzadas de análisis de sentimientos basadas en léxicos. Para obtener la información necesaria se recolectaron y analizaron documentos públicos de cinco destacadas startups de la región: Nu Bank, Merqueo, Platzi, La Haus y Rappi. Los resultados revelan que términos como "complejidad", "robo" y "fraude" predominan en el corpus de Nu Bank, indicando preocupaciones significativas sobre la usabilidad y la seguridad de los servicios. Se concluye que es esencial para las startups optimizar la experiencia del usuario, fortalecer las medidas de seguridad y mantener una comunicación transparente para gestionar eficazmente las expectativas de los inversionistas y reforzar la confianza de los usuarios. Este estudio ofrece valiosas perspectivas para otras startups en la región sobre la importancia de abordar estos desafíos críticos.

Palabras clave: análisis de sentimientos, startups, método basado en lexicones, minería de texto, expectativas de inversionistas, comunicaciones.

Abstract

Over the past decade, the proliferation of start-ups has seen significant growth, alongside advancements in technological innovation and access to funding in Latin America. This study explores the perceptions of start-up founders and managers regarding investor expectations using advanced lexicon-based sentiment analysis techniques. Public documents from five prominent start-ups in the region—Nu Bank, Merqueo, Platzi, La Haus, and Rappi—were collected and analyzed. The results reveal that terms such as "complexity," "theft," and "fraud" dominate Nu Bank's

Citación: Fortich Mesa, R. C. ., Muñoz Guzmán, M. ., & Santís Puche, M. A. . (2025). Análisis de sentimientos de startups usando un enfoque basado en léxicos. Publicaciones E Investigación, 19(1). https://doi.org/10.22490/25394088.9174

[1] docente_investigador3@uajs.edu.co - https://orcid.org/0000-0002-8076-1760
[2] docente_investigador4@uajs.edu.co - https://orcid.org/0000-0001-5624-6374
[3] direccion_administracion@uajs.edu.co - https://orcid.org/0000-0002-4366-7118

https://doi.10.22490/25394088.9174

corpus, indicating significant concerns about the usability and security of their services. It is concluded that it is essential for start-ups to optimize user experience, strengthen security measures, and maintain transparent communication to effectively manage investor expectations and reinforce user trust. This study offers valuable insights for other start-ups in the region on the importance of addressing these critical challenges.

Key words: *Sentiment analysis, startups, lexicon-based approach, text-mining, investor expectations, communications*

1. Introducción

En la última década, la proliferación de startups ha experimentado un notable crecimiento, paralelo a los avances en innovación tecnológica y acceso a financiación en América Latina. Este estudio explora las percepciones de los fundadores y gerentes de startups respecto a las expectativas de los inversionistas mediante técnicas avanzadas de análisis de sentimientos basadas en léxicos.

El capital de riesgo es una categoría de activos de importancia creciente: del total de empresas que se inscribieron en mercados de valores en Estados Unidos desde finales de los 1970, un 43 % tuvo apoyo de capital de riesgo previo a su OPA (Gornall & Strebulaev, 2015, como se citó en Gornall & Strebulaev, 2020).

Las startups ofrecen soluciones innovadoras pero sus fundadores y gerentes suelen enfrentar altas expectativas por parte de los inversionistas de capital de riesgo, sobre todo respecto a la valoración esperada en el mercado (Barg *et al*., 2021, Sievers *et al*., 2013; Seppä & Laamanen, 2001). Estas expectativas suelen ser muy exigentes, a veces con expectativas de multiplicaciones exponenciales de la valoración en menos de una década de iniciada la inversión, siendo crucial para los responsables de las startups comprender y gestionar estas expectativas adecuadamente.

La presente investigación tiene como objetivo responder la siguiente pregunta: ¿cómo es la opinión de los fundadores y gerentes de startups de América Latina sobre las expectativas de los inversionistas? Para abordar esta cuestión, se empleará una técnica de análisis de sentimientos basada en léxicos, con el fin de analizar y cuantificar las opiniones expresadas por los líderes de startups en la región.

El análisis de sentimientos es una técnica avanzada de procesamiento de lenguaje natural que se utiliza para determinar la polaridad (positiva o negativa) de un texto (Medhat *et al*., 2014; Mohammad, 2017; Behdenna *et al*., 2018; Zhao *et al*., 2016). Esta técnica es especialmente útil para analizar opiniones y sentimientos expresados en discursos, entrevistas, artículos y publicaciones en redes sociales. En el contexto de esta investigación, se adoptará un enfoque basado en léxicos para evaluar los sentimientos positivos y negativos presentes en las declaraciones de los fundadores y gerentes de startups respecto a las expectativas de los inversionistas. Este método es particularmente efectivo debido a su simplicidad y su capacidad para ofrecer resultados interpretables de manera expedita. La pregunta de investigación que se persigue responder con este estudio es: ¿cómo es la opinión de los fundadores y gerentes de startups de América Latina sobre las expectativas de los inversionistas?

2. Metodología

El análisis de sentimientos basados en diccionarios o lexicones es un método de evaluación de la polaridad de un texto usando conjuntos predefinidos de palabras (i.e. diccionarios) con sus correspondientes asociaciones sentimentales (Medhat *et al*., 2014). El sentimiento de un texto que siga este método debe comparar cada palabra con el diccionario para puntuar su polaridad individual y, al final, agregar la polaridad de todo el texto. En el enfoque básico bipolar (e.g. positivo y negativo) a

las palabras de un texto se les purga de aquellas que se consideren vacías y con las que se incluyan, llamadas palabras de opinión, se les cruza contra el diccionario para puntuarlas como positiva, si expresan algún estado deseado, o negativas, si expresan algún estado indeseado (Medhat *et al*., 2014; Nasukawa & Yi, 2003; Chong *et al*., 2014; Hogenboom *et al*., 2011).

El campo de análisis de sentimientos ha sido clasificado por Feldman (2013, p. 83) según la unidad de análisis en la que se concentra (i.e. documento, oración, aspecto, análisis comparativo de sentimiento y adquisición de diccionario de sentimiento). En este trabajo se aborda como unidad de análisis los documentos, y se acepta la premisa de que el documento contiene una opinión sobre una entidad, en general, o persona, evento o tema, en específico expresada por el autor del documento (Medhat *et al*., 2014).

El enfoque basado en diccionarios o lexicones de sentimientos se divide en: métodos basados en diccionarios, que parte de un conjunto de palabras semilla predefinidas que permiten expansión, y métodos basados en corpus, que comienzan con una lista inicial de palabras de opinión, denominada lista de palabras semilla y las cruza contra algún corpus grande buscando refinarla para algún contexto específico.

Como desventaja del método basado en diccionarios, es que a veces se torna en algo ultra-específico a un dominio en particular, en el que las palabras se asocien con una orientación positiva en un contexto común y corriente (e.g. el trigrama "prueba médica positiva"), pero que también se asocia con una orientación negativa en un contexto médico (e.g. significa que la persona tiene una enfermedad, lo que es negativo para su salud) (Wilson *et al*., 2009; Taboada *et al*., 2011; Choi & Cardie, 2008; Xia *et al*., 2016).

En este trabajo, se implementó un análisis de sentimientos con el paquete tidytext de R (Silge & Robinson, 2016). Este paquete contiene un conjunto de datos de sentimientos basado en diversos diccionarios. En particular, se empleó el diccionario NRC (Mohammad & Turney, 2013).

3. Participantes

Las empresas de la muestra para los datos de este trabajo fueron cinco empresas de tecnología o startups de América Latina con alto nivel de reconocimiento público: Nubank, Merqueo, Platzi, La Haus y Rappi (Tabla 1).

Tabla 1. Descripción de la muestra

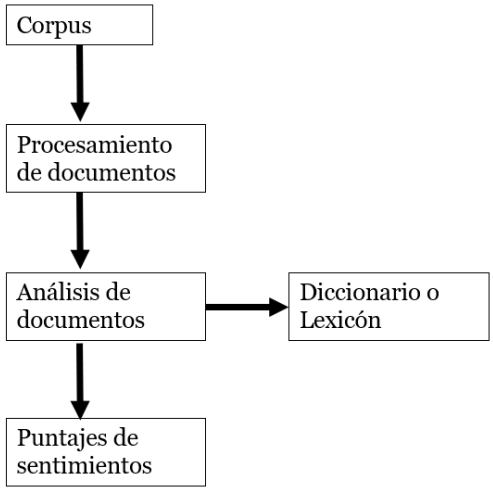
Nombre startup	País	Sector
Nubank	Brasil	Tecnología financiera
Merqueo	Colombia	Comercio electrónico
Platzi	Colombia	Educación en línea
La Haus	Colombia	Tecnología inmobiliaria
Rappi	Colombia	Entrega a domicilio

4. Resultados

Para la recolección de documentos, se descargaron de las páginas web de cada startup en estudio: comunicados de prensa, ensayos, columnas y blogs, y entrevistas con fundadores. Los criterios de inclusión usados fueron, primero, que estuvieran disponibles en línea; segundo, que estuvieran dirigidos directa o indirectamente a los inversionistas de las startups.

En el folder "Datos" se almacenaron tres archivos de texto con el texto crudo de las comunicaciones de cada startup, por ejemplo, *nu_01.txt* contiene una comunicación de Nu Bank. Luego, se usó una función para leer archivos de texto modificando la función genérica readLines() para que lea archivos con codificación UTF-8. Se prefirió usar readLines() por sobre las tradicionales read.table() o read.csv() porque el texto está sin estructura, y no se requiere que R asocie variables a columnas del archivo importado, solo se desea que el conjunto no estructurado de texto quede almacenado como objeto de R, y para eso, readLines() es la función idónea. La opción de encoding="UTF-8" permitió que los caracteres no ASCII puedan preservarse tras ser importados a R.

Figura 1. Sistema de análisis de sentimientos.



Nota. Adaptado de "Techniques and Applications for Sentiment Analysis: The Main Applications and Challenges of One of the Hottest Research Areas in Computer Science," por R. Feldman, 2013, *Communications of the ACM, 56*, p. 84. Derechos reservados [2013] por Association for Computing Machinery. Adaptado con permiso en trámite.

Un sistema de análisis de sentimientos suele componerse de cuatro etapas que, aunque parecen lineales, se apoyan durante la etapa del análisis de documentos de un diccionario o lexicón, retroalimentando la etapa de procesamiento de documentos (Figura 1). Se parte de un corpus de documentos como insumo del sistema, de ahí se anotan los documentos con anotaciones de sentimientos y finalmente se genera la variable puntajes de sentimientos.

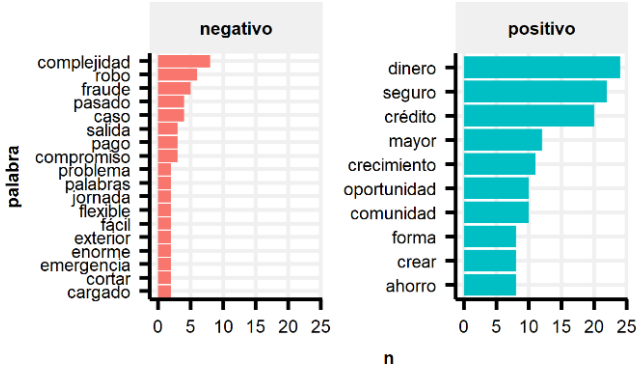
Algunos de los softwares más populares para la minería de texto son DiscoverText, AlchemyLanguage, WordStat, SAS Text Miner, MonkeyLearn, y R. Se eligió para este análisis R dada la conveniencia de su naturaleza de software libre.

Para calcular el puntaje de sentimientos se siguió un proceso de preprocesamiento del texto y análisis de sentimientos utilizando el paquete tidytext en R: primero, se cargaron y prepararon los datos. Inicialmente, se cargaron los paquetes necesarios y se prepararon los datos de las comunicaciones de las startups en un formato ordenado. Algunos de los paquetes cargados fueron: dplyr, stringr y tidytext.

Segundo, se unieron las versiones tokenizadas del texto con el diccionario de sentimientos. En R esto se logró mediante las herramientas del paquetev dplyr package, tales como la funcion inner join().

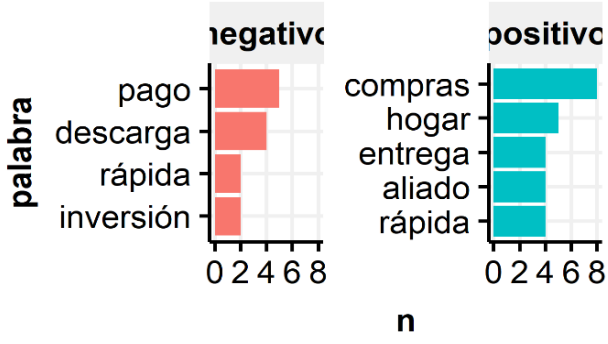
Finalmente, se calculó el puntaje de sentimientos como la frecuencia de menciones de términos detectados como negativos y positivos para un texto determinados, en comparación con el diccionario nrc.

Figura 2. Análisis de sentimientos Nu Bank.



El término negativo más prominente en el corpus de Nu Bank fue "complejidad" con una frecuencia de 5 a 10 repeticiones (Figura 1). Le siguió "robo" con una frecuencia ligeramente menor. En tercer lugar, de términos negativos frecuentes estuvo "fraude", con cerca de 5 repeticiones. Después, se halló que "pasado" siguió en frecuencia con cerca de 4 repeticiones en el corpus de esta startup. Las palabras positivas se concentraron en cambio en términos como "dinero" y "seguro".

Figura 3. Análisis de sentimientos Merqueo.



En el caso de Merqueo, la palabra con mayor frecuencia en el listado de sentimientos negativos fue "pago" con cerca de 5 ocurrencias, seguido de "descarga" y "rápida". Los sentimientos positivos aparecieron con términos como "compras" y "trabajo".

5. Discusión

Los responsables de Nu Bank mencionan en sus comunicaciones frecuentemente la complejidad como un aspecto que desean enfocar, posiblemente para convertirlo en un nicho de amigabilidad y menor complejidad que sus competidores.

El análisis de sentimientos aplicado a los documentos de Nu Bank ha proporcionado una visión detallada y cuantificable de la percepción pública y de los desafíos recurrentes que enfrentan los usuarios. A continuación, se detallan los hallazgos principales y su

interpretación en el contexto de Nu Bank y del ecosistema de startups en América Latina.

La prominencia de términos como "robo" y "fraude" es señal de preocupaciones importantes en la comunicación startup - público respecto a la seguridad de sus transacciones y datos personales. Este resultado puede complementarse con una revisión de recientes incidentes de seguridad en Nu Bank o rivales de mercado.

La prevalencia de términos negativos señala áreas críticas que requieren una atención inmediata. Estas áreas son esenciales para la satisfacción del cliente y la reputación de Nu Bank.

Aunque dominan los términos negativos, también se identificaron términos positivos respecto de los servicios de las startups, tales como la seguridad financiera y la confiabilidad en determinadas operaciones.

Las startups deberían comunicar claramente las soluciones implementadas y los avances realizados, para así mitigar las percepciones negativas persistentes y fortalecer la relación con los usuarios.

Los hallazgos de este análisis proporcionan valiosas lecciones para otras startups en América Latina. Aspectos críticos como la simplicidad de uso, la robustez de las medidas de seguridad, y la transparencia en la comunicación son factores determinantes para el éxito en un mercado tecnológico altamente competitivo.

Este estudio tuvo la limitación de que se basó en documentos disponibles públicamente, lo que puede no reflejar todos los aspectos de las empresas. También, el enfoque basado en diccionarios tiene limitaciones intrínsecas, tales como la incapacidad de capturar matices contextuales o identificar sarcasmo e ironía en los textos y se puede complementar con otros enfoques de aprendizaje de máquinas.

Referencias

Barg, J. A., Drobetz, W. & Momtaz, P. P. (2021). Valuing start-up firms: A reverse-engineering approach for fair-value multiples from venture capital transactions. *Finance Research Letters*, *43*, 102008. https://doi.org/https://doi.org/10.1016/j.frl.2021.102008

Behdenna, S., Barigou, F. & Belalem, G. (2018). Document level sentiment analysis: A survey. *EAI Endorsed Transactions on Context-aware Systems and Applications*, *4*(13), e2. https://doi.org/10.4108/eai.14-3-2018.154339

Caparroso, J. (2021). Merqueo lanza entregas de super en menos de 10 minutos en la cdmx. *Forbes México*. https://www.forbes.com.mx/merqueo-lanza-entregas-de-super-en-menos-de-10-minutos-en-la-cdmx/

Choi, Y. & Cardie, C. (2008). Learning with compositional semantics as structural inference for subsentential sentiment analysis. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 793–801.

Chong, W. Y., Selvaretnam, B. & Soon, L.-K. (2014). Natural language processing for sentiment analysis: An exploratory analysis on tweets. *2014 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology*, 212–217. https://doi.org/10.1109/ICAIET.2014.43

Feldman, R. (2013). Techniques and applications for sentiment analysis: The main applications and challenges of one of the hottest research areas in computer science. *Communications of the ACM*, *56*(4), 82–89. https://doi.org/10.1145/2436256.2436274

Gornall, W. & Strebulaev, I. A. (2020). Squaring venture capital valuations with reality. *Journal of Financial Economics, 135*(1), 120-143.

Hogenboom, A., van Iterson, P., Heerschop, B., Frasincar, F. & Kaymak, U. (2011). Determining negation scope and strength in sentiment analysis. *2011 IEEE International Conference on Systems, Man, and Cybernetics*, 2589–2594. https://doi.org/10.1109/ICSMC.2011.6084066

Medhat, W., Hassan, A. & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal, 5*(4), 1093–1113. https://doi.org/https://doi.org/10.1016/j.asej.2014.04.011

Mohammad, S. M. (2017). Challenges in sentiment analysis. In E. Cambria, D. Das, S. Bandyopadhyay & A. Feraco (Eds.). *A practical guide to sentiment analysis* (pp. 61–83). Springer International Publishing. https://doi.org/10.1007/978-3-319-55394-8 4

Mohammad, S. M. & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence, 29*(3), 436–465. https://doi.org/10.1111/j.1467-8640.2012.00460.x

Nasukawa, T., & Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. *Proceedings of the 2nd International Conference on Knowledge Capture*, 70–77.

Sadia, A., Khan, F. & Bashir, F. (2018). An overview of lexicon-based approach for sentiment analysis. *2018 3rd International Electrical Engineering Conference* (IEEC 2018), 1–6.

Sievers, S., Mokwa, C. F. & Keienburg, G. (2013). The relevance of financial versus non-financial information for the valuation of venture capital-backed firms. *European Accounting Review, 22*(3), 467–511. https://doi.org/10.1080/09638180.2012.741051

Silge, J. & Robinson, D. (2016). Tidytext: Text mining and analysis using tidy data principles in r. *JOSS, 1*(3). https://doi.org/10.21105/joss.00037

Taboada, M., Brooke, J., Tofiloski, M., Voll, K. & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics, 37*(2), 267–307. https://doi.org/10.1162/COLI a 00049

Wilson, T., Wiebe, J. & Hoffmann, P. (2009). Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics, 35(*3), 399–433.

Xia, R., Xu, F., Yu, J., Qi, Y. & Cambria, E. (2016). Polarity shift detection, elimination and ensemble: A three-stage model for document-level sentiment analysis. *Information Processing & Management, 52*(1), 36-45. https://doi.org/https://doi.org/10.1016/j.ipm.2015.04.003

Zhao, J., Liu, K. & Xu, L. (2016). Sentiment analysis: Mining opinions, sentiments, and emotions. *Computational Linguistics, 42*(3), 595-598. https://doi.org/10.1162/COLI r 00259

Zirn, C., Niepert, M., Stuckenschmidt, H. & Strube, M. (2011). Fine-grained sentiment analysis with structural features. *Proceedings of 5th International Joint Conference on Natural Language Processing*, 336–344.